Appendix: Preliminary Theoretical Results on the Convergence of MISSION

Amirali Aghazadeh [*1] **Ryan Spring** [*2] **Daniel LeJeune** [3] **Gautam Dasarathy** [3] **Anshumali Shrivastava** [2] **Richard G. Baraniuk** [3]

In this appendix, we present some preliminary results on the convergence of **MISSION**. For the sake of exposition, we will consider the full-gradient descent version of **MISSION**, and we will prove that the iterates converge geometrically upto a small additive error. In order to establish this proof, we make an assumption (Assumption 1) about the hashing scheme; see Section 0.1 for more on this.

We begin by establishing some notation. We will assume that the data satisfies the following linear model:

$$y = X\beta^* + w, \tag{1}$$

where $y \in \mathbb{R}^n$ is the vector of observation, $X \in \mathbb{R}^{n\times p}$ is the data matrix, $w \in \mathbb{R}^n$ is the noise vector, and $\beta^* \in \mathbb{R}^p$ is the unknown $k-$sparse regression vector. We will let ψ and φ respectively denote the hashing and the (top-k) heavy-hitters operation. We will let β^t denote the output of **MISSION** in step t. In general, we will let the vector $h \in \mathbb{R}^m$ denote the hash table. Finally, as before, we will let H_k denote the projection operation onto the set of all $k-$sparse vectors. We will make the following assumption about the hashing mechanism:

Assumption 1. *For any $h \in \mathbb{R}^m$, there exists an $\beta_h \in \mathbb{R}^p$ such that the following hold*

1. *$\psi(\beta_h) = h$, that is, the hash table contents can be set to h by hashing the vector β_h.*
2. *$\|\beta_h - H_k(\beta_h)\|_2 \le \varepsilon_1$*

This assumption requires the hashing algorithm to be such that there exists a nearly sparse vector that can reproduce any state of the hash table exactly. This is reasonable since the hash table is a near optimal "code" for sparse vectors in $\mathbb{R}^p$. See Section 0.1 for more on this.

[*]Equal contribution [1]Department of Electrical Engineering, Stanford University, Stanford, California [2]Department of Computer Science, Rice University, Houston, Texas [3]Department of Electrical and Computer Engineering, Rice University, Houston, Texas. Correspondence to: Anshumali Shrivastava <anshumali@rice.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

We will next state a straightforward lemma about the sketching procedure

Lemma 1. *There exist constants $\varepsilon_2, C_1 > 0$ such that provided that the size m of the hash table satisfies $m \ge C_1 k \log^2 p$, the following holds for any $\beta \in \mathbb{R}^p$ with probability at least $1 - \delta_1$:*

$$\|\varphi(\psi(\beta)) - H_k(\beta)\|_2 \le \varepsilon_2 \tag{2}$$

This lemma follows directly from the definition of the Count-Sketch, and we will not prove here.

We next state the main theorem that we will show.

Theorem 1. *For any $\delta \in \left(0, \frac{1}{3}\right)$ and $\rho \in (0, 0.5)$, there is a constant $C > 0$ such that the following statement holds with probability at least $1 - 3\delta$*

$$\left\|\beta^{t+1} - \beta^*\right\|_2 \le 2\rho \left\|\beta^t - \beta^*\right\|_2 + 2\sqrt{\frac{\sigma_w^2(1+\mu)k\log p}{n}} + 2\varepsilon_1 + 3\varepsilon_2, \tag{3}$$

provided that $n > Ck\log p$, $m > Ck\log^2 p$, and that Assumption 1 holds.

Notice that since $\rho < 0.5$, the above theorem guarantees geometric convergence. This implies that the overall error is of the order of the additive constants ε_1 and ε_2.

Before we prove this theorem, we will collect some lemmas that will help us prove our result.

Lemma 2. *Suppose $X \in \mathbb{R}^{n\times p}$ has i.i.d $\mathcal{N}(0, \frac{1}{n})$ entries. Then for constants $\rho, \delta_2 > 0$, there exists a constant $C_2(\delta) > 0$ such that if $n \ge C_2 k \log p$ such that for any pair of unit-norm $k-$sparse vectors $\beta_1, \beta_2 \in \mathbb{S}^{p-1}$, the following holds with probability at least $1 - \delta_2$.*

$$|\langle X\beta_1, X\beta_2\rangle - \langle \beta_1, \beta_2\rangle| \le \rho. \tag{4}$$

Proof. Note that $\mathbb{E}[\langle X\beta_1, X\beta_2\rangle] = \langle \beta_1, \beta_2\rangle$. For a fixed pair of β_1, β_2, the proof follows from a standard Chernoff bound argument after observing that $\langle X\beta_1, X\beta_2\rangle$ can be written as a sum of products of independent Gaussian random variables. The rest of the proof follows from a standard covering argument, which gives the requirement on n. □

Lemma 3. *Suppose X has i.i.d entries drawn according to $\mathcal{N}(0, n^{-1})$, and $w \sim \mathcal{N}(0, \sigma_w^2 I_n)$ is drawn independently of X. Then, for any constant $\delta_3 > 0$, there are constants $C_3, \mu > 0$ such that for all unit norm $k-$sparse $\beta \in \mathbb{S}^{p-1}$, the following holds with probability at least $1 - \delta_3$:*

$$\langle \beta, X^T w \rangle \leq \sqrt{\frac{\sigma_w^2(1+\mu)k \log p}{n}} \tag{5}$$

provided $n \geq C_3 k \log p$.

Proof. Notice that for a fixed β, $\langle \beta, X^T w \rangle = \langle X\beta, w \rangle$ has the same distribution as $\frac{1}{\sqrt{n}} \|w\|_2 \langle \beta, w_2 \rangle$, where $w_2 \sim \mathcal{N}(0, I_n)$ is independent of w. Now, we can use concentration inequalities of chi-squared random variables to show that there is a constant $C_3' > 0$

$$\mathbb{P}\left[\|w\|_2^2 \geq \sigma_w^2(1+\mu_1)n\right] \leq e^{-C_3' n}. \tag{6}$$

Similarly, from chi-squared concentration, there is a constant $C_3'' > 0$

$$\mathbb{P}\left[|\langle \beta, w_2 \rangle|^2 \geq 1 + \mu_2\right] \leq e^{-C_3''} \tag{7}$$

Now, with a standard covering argument, we know that there is a constant $C_3''' > 0$ such that provided $n > C_3''' k \log p$, the following holds for at least $1 - \delta_3$ for any $k-$sparse β:

$$\begin{aligned}\langle \beta, A^T w \rangle &= \langle A\beta, w \rangle \\ &\leq \sqrt{\frac{\sigma_w^2(1+\mu)nk \log p}{n}}.\end{aligned}$$

□

Proof of Theorem 1 If we let h^t denote the contents of the hash table at round t, notice that we have the following: $x^{t+1} = \varphi(h^{t+1})$. The (full gradient descent version of the) **MISSION** algorithm proceeds by updating the hash table with hashes of the gradient updates. Therefore, we have the following relationship:

$$h^{t+1} = h^t + \psi\left(\eta X^T X(\beta^* - \beta^t) + X^T w\right), \tag{8}$$

where β^t is the output of the algorithm at round t. Notice that $\beta^t = \varphi(h^t)$. According to Assumption 1, we know that there exists a vector $\tilde{\beta}^t$ such that $\psi(\tilde{\beta}^t) = h^t$. We will use this observation next. Notice that the output of round $t+1$ maybe written as follows:

$$\begin{aligned}\beta^{t+1} &= \varphi\left(h^t + \psi\left(\eta X^T X(\beta^* - \beta^t) + X^T w\right)\right) \\ &= \varphi\left(\psi\left(\tilde{\beta}^t + \eta X^T X(\beta^* - \beta^t) + X^T w\right)\right).\end{aligned}$$

Now, we will estimate how close the output of the algorithm gets to β^* in round $t+1$ in terms of how close the algorithm got in round t. Notice that

$$\begin{aligned}&\left\|\beta^{t+1} - \beta^*\right\|_2 \\ &= \left\|\varphi\left(\psi\left(\tilde{\beta}^t + \eta X^T X(\beta^* - \beta^t) + X^T w\right)\right) - \beta^*\right\|_2 \\ &\leq \left\|H_k\left(\tilde{\beta}^t + \eta X^T X(\beta^* - \beta^t) + X^T w\right) - \beta^*\right\|_2 + \varepsilon_2,\end{aligned} \tag{9}$$

which follows from Lemma 1. We will next consider the first term from above. For notational ease, we will set $\gamma^{t+1} \triangleq \tilde{\beta}^t + \eta X^T X(\beta^* - \beta^t) + X^T w$. Observe that H_k is an orthogonal projection operator, and that β^* is $k-$sparse, therefore we have that

$$\left\|H_k\left(\gamma^{t+1}\right) - \gamma^{t+1}\right\|_2^2 \leq \left\|\gamma^{t+1} - \beta^*\right\|_2^2. \tag{10}$$

Adding and subtracting β^* on the left side and cancelling out the common terms, we have the following.

$$\begin{aligned}&\left\|H_k(\gamma^{t+1}) - \beta^*\right\|_2^2 \\ &\leq 2\langle H_k(\gamma^{t+1}) - \beta^*, \gamma^{t+1} - \beta^* \rangle \\ &= 2\langle H_k(\gamma^{t+1}) - \beta^*, \tilde{\beta}^t + \eta X^T X(\beta^* - \beta^t) + X^T w - \beta^* \rangle \\ &= 2\langle H_k(\gamma^{t+1}) - \beta^*, \beta^t + \eta X^T X(\beta^* - \beta^t) + X^T w - \beta^* \rangle \\ &\qquad + 2\langle H_k(\gamma^{t+1}) - \beta^*, \beta^t - \tilde{\beta}^t \rangle \\ &\overset{(a)}{\leq} 2\langle H_k(\gamma^{t+1}) - \beta^*, \beta^t + \eta X^T X(\beta^* - \beta^t) + X^T w - \beta^* \rangle \\ &\qquad + 2\left\|H_k(\gamma^{t+1}) - \beta^*\right\|_2 \left\|\varphi(\psi(\tilde{\beta}^t)) - \tilde{\beta}^t\right\|_2 \\ &\leq 2\langle H_k(\gamma^{t+1}) - \beta^*, \beta^t + \eta X^T X(\beta^* - \beta^t) + X^T w - \beta^* \rangle \\ &\qquad + 2\left\|H_k(\gamma^{t+1}) - \beta^*\right\|_2 \Big(\left\|H_k(\tilde{\beta}^t) - \tilde{\beta}^t\right\|_2 + \\ &\qquad\qquad \left\|H_k(\tilde{\beta}^t) - \varphi(\psi(\tilde{\beta}^t))\right\|_2\Big) \\ &\overset{(b)}{\leq} 2\langle H_k(\gamma^{t+1}) - \beta^*, \beta^t + \eta X^T X(\beta^* - \beta^t) + X^T w - \beta^* \rangle \\ &\qquad + 2\left\|H_k(\gamma^{t+1}) - \beta^*\right\|_2 (\varepsilon_1 + \varepsilon_2),\end{aligned} \tag{11}$$

where (a) follows form the Cauchy-Schwarz inequality and from the definition of $\tilde{\beta}^t$, (b) follows from Assumption 1 and Lemma 1. We will now turn our attention to the first inner-product in (11). With some rearrangement of terms,

one can see that

$$\begin{aligned}
&\langle H_k(\gamma^{t+1}) - \beta^*, \beta^t + \eta X^T X(\beta^* - \beta^t) + X^T w - \beta^* \rangle \\
&= \langle H_k(\gamma^{t+1}) - \beta^*, \beta^t - \beta^* \rangle - \eta \langle X \left(H_k(\gamma^{t+1}) - \beta^* \right), \\
&\qquad X(\beta^t - \beta^*) \rangle + \eta \langle H_k(\gamma^{t+1}) - \beta^*, X^T w \rangle \\
&\overset{(a)}{\leq} \rho \left\| H_k(\gamma^{t+1}) - \beta^* \right\|_2 \left\| \beta^t - \beta^* \right\|_2 \\
&\qquad + \langle H_k(\gamma^{t+1}) - \beta^*, X^T w \rangle \\
&\overset{(b)}{\leq} \rho \left\| H_k(\gamma^{t+1}) - \beta^* \right\|_2 \left\| \beta^t - \beta^* \right\|_2 \\
&\qquad + \left\| H_k(\gamma^{t+1}) - \beta^* \right\|_2 \sqrt{\frac{\sigma_w^2 (1+\mu) k \log p}{n}}
\end{aligned} \tag{12}$$

where (a) follows from Lemma 2 and setting $\eta = 1$. (b) follows from Lemma 3.

Putting (11) and (12), we get

$$\begin{aligned}
\left\| H_k\left(\gamma^{t+1}\right) - \gamma^t \right\|_2 &\leq 2\rho \left\| \beta^t - \beta^* \right\|_2 \\
&\quad + 2\sqrt{\frac{\sigma_w^2 (1+\mu) k \log p}{n}} + 2\left(\varepsilon_1 + \varepsilon_2\right).
\end{aligned} \tag{13}$$

Putting this together with (9) gives us the desired result.

0.1. On Assumption 1

In the full-gradient version of the **MISSION** algorithm, one might modify the algorithm explicitly to ensure that Assumption 1. Towards this end, one would simply ensure that the gradients vector is attenuated on all but its top k entries at each step. It is not hard to see that this *clean-up* step will ensure that Assumption 1 holds and the rest of the proof simply goes through.

In **MISSION** as presented in the manuscript, we employ stochastic gradient descent (SGD). While the above proof needs to be modified for it to be applicable to this case, our simulations suggest that this clean-up step is unnecessary here. We suspect that this is due to random cancellations that are introduced by the SGD. This is indeed an exciting avenue for future work.